Post-Prediction Inference on Political Twitter

Alicia Gunawan, Dylan Haar, Luis Ledezma-ramos

1 Intro

Machine learning is a modern task in data science that uses observed data values to model and predict data. It takes advantage of having observed data available, but what should be done when observed data cannot be obtained? A common practice is to use predicted values when observed values are unavailable, but without any corrections we inevitably run into issues such as deflated standard errors, bias, and inflated false positive rates.

Wang et al. proposes a method to correct inference done on predicted outcomes–which they name post-prediction inference, or postpi–in *Methods for correcting inference based on outcomes predicted by machine learning*. This statistical technique takes advantage of the standard structure for machine learning and uses bootstrapping to correct statistical error using predicted values in place of observed values.

We are exploring the applicability of Wang et al.'s postpi bootstrapping technique on political data–that is, on political twitter posts. Our project will be investigating what kinds of phrases or words in a tweet will strongly indicate a person's political alignment, in the context of US politics. By doing so, we can simultaneously test how the bootstrap post-prediction inference approach interacts with Natural Language Processing models and how this method can be generally applicable towards analyses in political science.

2 Methodology

The postpi bootstrap approach by Wang et al. is a method that aims to correct inference in studies that use predicted outcomes in lieu of observed outcomes. It is effective due to its simplicity–this approach is not dependent on deriving the first principles of the prediction model, so we are free to focus on accuracy without worrying about the impacts towards the complexity of the model. The reason why it is not dependent is because this approach utilizes an easily generalizable and low-dimensional relationship between observed and predicted outcomes.

An implementation for this algorithm is provided below:

Algorithm Bootstrap-based Postpi Correction

Require: Observations $\{x_{(tr)}, y_{(tr)}, x_{(te)}, y_{(te)}, x_{(val)}\}$, where $y \in \mathbb{R}^n$ and $x \in \mathbb{R}^{n \times m}$.
Require: Prediction model $\hat{f}(\cdot)$, relationship model $k(\cdot)$, and inference model of the form $g[E(y|X)] = X\beta$.
1: Use $(x_{(tr)}, y_{(tr)})$ to fit the prediction model s.t. $y_p = \hat{f}(x)$.
2: Get test set predicted outcomes: $y_{p(te)} = \hat{f}(x_{(te)})$.
3: Use $(y_{(te)}, y_{p(te)})$ to fit the relationship model s.t. $y = k(y_p)$.
4: Get validation set predicted outcomes: $y_{p(val)} = \hat{f}(x_{(val)})$
5: Use $(x_{(val)}, y_{p(val)})$ to bootstrap
6: **for** $b \in \{1, ..., B\}$ **do**
7: Sample predicted outcomes and covariates $(x^b_{i(val)}, y^b_{pi(val)})$ with replacement for $i = 1, ..., n$.
8: Simulate values from the relationship model $\bar{y}^b_i = k(y^b_{p(val)})$.
9: Fit the inference model $g[E(\bar{y}^b|X^b_{(val)})] = X^b_{(val)}\beta^b$.
10: Extract coefficient estimator β^b from the fitted inference model.
11: Extract the SE of the estimator $se(\hat{\beta}^b)$ from the fitted inference model.
12: Estimate the inference model coefficient using a median function $\hat{\beta}^{boot} = median(\hat{\beta}^1, \hat{\beta}^2, ..., \hat{\beta}^B)$.
13: Estimate the inference model SE:
14: The parametric method $\hat{SE}^{boot,par} = median(\hat{SE}(\hat{\beta}^1), \hat{SE}(\hat{\beta}^2), ..., \hat{SE}(\hat{\beta}^B))$.
15: The nonparametric method $\hat{SE}^{boot,non-par} = SD(\hat{\beta}^1, \hat{\beta}^2, ..., \hat{\beta}^B)$.

3 Data

3.1 Data Collection

We collected our data by scraping tweets from US politicians from Twitter. Specifically, we took the Twitter handles of the President, Vice President, and all the members of US Congress except Representatives Chris Smith (R-NJ) and Jefferson Van Drew (R-NJ), as they have both deleted their Twitter accounts. These Twitter handles were compiled and provided by the UCSD library, and outdated names or Twitter handles were updated manually by ourselves. Additionally, the two Independent members of Congress–Sen. Bernie Sanders (I-VT) and Angus King (I-ME)–will be considered Democratic politicians for our purposes, as they caucus with Democrats.

Using these Twitter handles, we scraped approximately 100 tweets from each politician, although the exact number of tweets pulled from each individual will fluctuate as not all members of Congress use Twitter with the same frequency as their colleagues. Our final dataset consists of 44,328 tweets for an average of 82 tweets per politician. Of these tweets, 22,653 tweets are from Democrats, 21,478 tweets are from Republicans, and 197 tweets are from Independents (converted to Democrats).

3.2 Cleaning

To prepare our data for prediction and feature selection, we cleaned the tweets by expanding all contractions, moved all text into lowercase format, and removed urls, punctuation, and unicode

characters. Additionally, we also removed stopwords, using the dictionary of stopwords provided by NLTK to do so.

3.3 Exploratory Data Analysis

Our data consists of a relatively equal number of tweets leaning either Democratic or Republican. As said earlier, with Independent politicians counting as Democrats, there are a total of 44,328 tweets–22,850 are classified as tweets from Democrats, while 21,478 are classified as tweets from Republicans.

We look at Figure 1 for a first glance at the data. Figure 1 is an overlaid histogram plotting the number of words in tweets from Democrats and Republicans. While both histograms are clearly skewed to the left, we can see that the distribution of the length of tweets for Democrats

The distribution of the length of tweets for Republicans clearly skews more left than the distribution for Democrats, which tells us that tweets from Republicans average less words compared to their counterparts on the opposite aisle. This could imply that the prediction model will utilize more vocabulary from Democrat-classified tweets than Republican, which might have interesting effects on the prediction model and thus the bootstrap algorithm and inference.

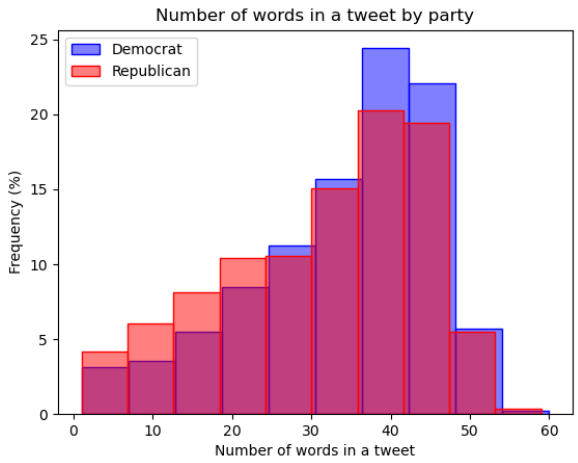

Figure 1: A histogram depicting the number of words in a tweet by party. We can see that Democrats generally have longer tweets compared to Republicans.

We take a deeper dive into each classification in Figure 2 below, which lists the 10 most frequent words used by each party, excluding stopwords. There are very few commonalities between either party–only two words are commonly used by both parties: ‘today’ and ‘year’.

Democrats seem to focus on policy issues as suggested by ‘act’ and ‘infrastructure’, but otherwise their attentions are spread across a multitude of topics as no single unifying issue seems to be able to group together their most frequently used words. On the other hand,

Republicans seem to focus more on their political opponents–words such as ‘biden’, ‘democrats’, and ‘president’ seem to suggest that–and on the American people. There is notably a significant reference to ‘biden’, with it being used approximately 3500 times, almost double the frequency of the second most popular word.

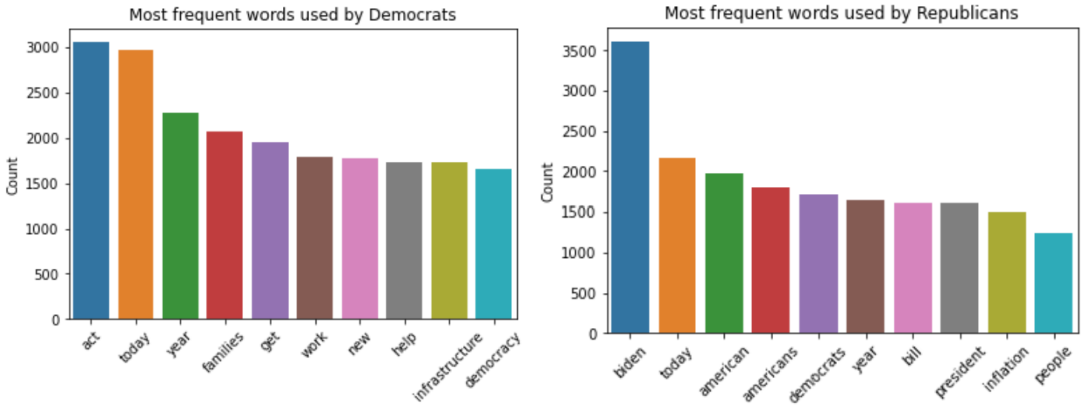

Figure 2: Bar plots depicting the most frequent words used by either party. We can also see a significant difference in the most frequent words used by either party–only ‘today’ and ‘year’ is a word that both parties use in common.

4 Methods

4.1 Prediction and Relationship Model

During this stage of our project, we worked on maximizing the accuracy of our prediction model. In order to do so, we used a TF-IDF vectorization model with 200,000 features and 1-3 words per feature, and an SVC model for prediction, with a linear kernel and C=1.5.

We compared several different prediction models in the process of coming up with our final SVC model. These included other classification algorithms such as logistic regression and ridge regression (regularized). After determining which model performed the best, we tuned hyperparameters on that model to further improve its performance.

Following the method that Wang et al. used to prepare the prediction data for the bootstrap postpi method, we used our prediction model to generate the probability distribution for each tweet–the probability of it being classified as Democratic, Republican, or Independent-leaning–and used this data and the observed outcomes from the test dataset to build a relationship model. We used a K-NN machine learning model for this.

4.2 Feature Selection for Inference

We reviewed relevant literature in political science to develop a criteria for choosing our features.

In *Twitter Language Use Reflects Psychological Differences between Democrats and Republicans*, Sylwester and Pulver discuss the differences between Democrats and Republicans in the context of previous findings and their own discoveries. For example, Haidt's Moral Foundations model, which identifies "harm, fairness, liberty, ingroup, authority, and purity" as the pillars of morality, has been used to distinguish between liberals and conservatives. It was found that liberals prioritized the harm and fairness aspects of morality, while conservatives focused more on liberty, ingroup, authority, and purity. Sylwester and Pulver also found differences between Democratic and Republican-aligned people when it came to what kinds of topics they discussed and emotions they expressed. Republicans focused more on topics such as "religion…, national identity…, government and law…, and their opponents" while Democrats were focused on emphasizing their uniqueness and generally expressed more anxiety and emotion.

We also reviewed Chen et al.'s study, *#Election2020: the first public Twitter dataset on the 2020 US Presidential election*. Chen et al. found that more conservative Twitter users tended to share more topics related to conspiracy theories and "public health and voting misinformation" compared to liberal Twitter users.

Taking these two sources into consideration, our criteria for selecting features was whether or not they would fall into either liberal or conservative tendencies as discovered by either source. If a feature implied a discussion of harm or fairness, or was an expression of uniqueness, anxiety, or emotion, then we anticipated that this feature would connect more to Democratic-aligned tweets. On the other hand, if a feature discussed liberty, purity, religion, national identity, government and law, or Republican opponents, or implied that the topic at hand was associated with public health or voting misinformation, said feature may be connected to Republican-aligned tweets.

We ended up choosing 5 features to conduct inference, which are border, illegals, god, defund, and love. We hypothesized that the first three would be strong indicators for a Republican-classified tweet as they allude to national identity and religion, while the last two would indicate a Democratic-classified tweet as they allude to concepts of harm and fairness, as well as emotion.

5 Results

After conducting inference using the bootstrap postpi algorithm, we found that the parametric method worked best to correct for inference.

Interpretation of results

- Nonparametric is inferior in all regards
    - SE = SD of estimators, low SD of predicted values tells us that the estimators sampled by bootstrapping are clustered around the mean

        - Bootstrapping prediction function captures the mean, but doesn't capture the variance?
    - In Wang et al. non-parametric also consistently is inferior to parametric method

6 Conclusion

<blank>

7 References

Chen, E., Deb, A. & Ferrara, E. #Election2020: the first public Twitter dataset on the 2020 US Presidential election. *J Comput Soc Sc* (2021). https://doi.org/10.1007/s42001-021-00117-9

Sylwester K, Purver M (2015) Twitter Language Use Reflects Psychological Differences between Democrats and Republicans. PLOS ONE 10(9): e0137422. https://doi.org/10.1371/journal.pone.0137422

Wang, Siruo, Tyler H. McCormick, and Jeffrey T. Leek. "Methods for correcting inference based on outcomes predicted by machine learning." *Proceedings of the National Academy of Sciences* 117.48 (2020): 30266-30275.